Back to the drawing board: Can we compare socioeconomic background scales?[1]

Andrés Sandoval-Hernandez
University of Bath
David Rutkowski
Indiana University
Tyler Matta
Pearson
University of Oslo, Centre for Educational Measurement
Daniel Mirnda
Pontificia Universidad Católica de Chile, MIDE UC
Centro de Estudios de Conflicto y Cohesión Social

Abstract

Using data from international large-scale assessments (ILSA), we evaluate the issue of country-level model-data consistency of background socio-economic scales, as well as the invariance across countries. To that end, we use data from PISA, TERCE, and TIMSS, as they operationalize socio-economic status somewhat differently. As part of our analysis, we examine whether TERCE, a Latin American study – with measures that are regionally developed – exhibits better psychometric properties than measures that are designed to function across a larger and more diverse number of educational systems. We also examine TIMSS, a trends focused study – that has historically emphasized consistency and comparison. Finally, we include PISA which has the largest number of participants and has changed and conceptualized a great deal of its background questionnaire depending on the study's major domain and focus. Our findings suggest that none of the socio-economic background scales we analyzed are fully invariant in any of the three studies, and therefore comparisons across countries should be done with caution. The different levels of equivalence reached by each scale in each study and the type of comparisons that can be made given these results (e.g., comparison of average scale scores, comparison of relationships between the tested scales and other variables) are discussed in the full paper.

Key words: measurement invariance, measurement equivalence, TERCE, TIMSS, PISA, multi-group confirmatory factor analysis, socioeconomic scales.

Introduction

International large-scale assessments (ILSAs) of educational achievement such as the Programme for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), and the Third Regional Comparative and Explanatory Study (TERCE) serve multifold purposes. From system monitoring and benchmarking to providing national participants and researchers with information about what students know and can do. ILSAs offer stakeholders an opportunity for understanding the context and correlates of learning in a number of areas as well as provide important background information on students, teachers, and schools. As national participation in these assessments grows, however, it is becoming more difficult for testing organizations to tailor assessments to meet the needs of a diverse set of participants. For example, in the 2015 PISA cycle, all 34 OECD member countries (representing the largest economies in the world, excepting China) participated in PISA, with the remaining 38 participants (termed *partner* systems) comprised of a heterogeneous mix of economies and cultures, including educational systems such as Tunisia, Peru, Singapore, and Shanghai, China. A similar situation is also faced in

[1] This paper was supported by the Norwegian Research Council Grant #255246 and by the Center of Social Conflict and Cohesion Studies —COES CONICYT/FONDAP N°°15130009

TIMSS. Finally, although regional assessments such as TERCE have fewer participants and – ostensibly – less heterogeneity than more global international studies, language, economic and cultural diversity among and within TERCE participants persists. For example, Chile's GDP per capita is over three times that of Bolivia and although the majority of countries share Spanish as a common language many participating countries include indigenous populations with a variety of mother tongues.

Most ILSAs include both a cognitive assessment and a set of background questionnaires. The questionnaires are administered to students and, depending on the assessment, can measure others such as teachers, parents and school leaders. In general, the background questionnaires have two primary uses: (1) to help contextualize the assessed educational system; and (2) to optimize population and sub-population achievement estimation. The benefits of using background data to help estimate achievement are well documented (Mislevy, Beaton, Kaplan, and Sheehan, 1992) and are not the focus of this paper. In fact, potential methodological challenges associated with an amalgamation of participants have been highlighted by a number of researchers, who have pointed especially to the achievement estimation model and whether comparisons are sensible and valid when systems differ dramatically (Goldstein, 2004; Kreiner and Christensen, 2014; Mazzeo and von Davier, 2009; Oliveri and Ercikan, 2011). Partly in response to these and other criticisms, the PISA project has implemented accommodations, especially targeted toward lower performing participants (e.g., incorporating easier items into the test for countries with low expected performance; OECD, 2012). Recent research has demonstrated that these types of accommodations are a promising way of acknowledging and dealing with the heterogeneity that is necessarily present in cross-cultural research (Rutkowski, Rutkowski and Zhou, 2016). Significant work has been done to ensure comparability of achievement scales across countries (e.g. OECD, 2014; Schulz, Ainley and Frailon, 2011; UNESCO-OREALC, 2016) and across time (e.g. Gaviria and Covadonga, 2007). In contrast, much less effort is spent on designing scales derived from the background questionnaires that can account for vast differences among participants (Rutkowski and Rutkowski, 2010).

Empirically, research has shown that the assumption of equivalent background scales in ILSAs is often violated, leading to compromised comparability (Caro, Sandoval-Hernandez and Lüdtke, 2016; Glas and Jehangir, 2014; Oliveri and von Davier, 2014). As such, the objective of this paper is twofold: First, to demonstrate a method to explore both within-county data consistency and equivalence across countries on background scales. Second, to discuss the results of the application of this method to the socio-economic scales of PISA, TIMSS and TERCE. More specifically, we explore the different levels of equivalence reached by the scales used in each study to measure some form of socio-economic status (SES) and discuss the type of comparisons that can be made given these results (e.g., comparison of average scale scores, comparison of relationships between the tested scales and other variables).

It would not be feasible to evaluate the equivalence of all the background scales of the three studies, for this reason in this paper we focus on the scales developed by the testing organizations to examine some form of SES in three international studies (PISA, TERCE, and TIMSS). We decided to use these scales because, in the studies focused on identifying factors associated with learning outcomes (e.g. school and teacher effectiveness), SES is the control variable that consistently shows a stronger association with educational achievement. Furthermore, there is an important body of literature specifically focused on understanding the mechanisms by which socio-economic background or family socio-economic status is associated with academic achievement (Buchmann, 2002).

By examining the equivalence of these scales between countries and comparing the findings across studies, we can determine if different assessment designs or approaches result in different degrees of comparability. The three studies were purposely chosen because they represent three different designs of international assessment. TERCE was chosen to represent a regional study – with measures that are regionally developed and with the assumption that test developers were able to focus the scale for a smaller group participants (Treviño, Fraser, Meyer, Morawietz, Inostroza and Naranjo, 2015). TIMSS was chosen to represent a trend focused study – one that has

historically emphasized consistency and comparison over changes in societies, constructs, or participants. Finally, we include PISA, which has the largest number of participants and has historically been willing to make significant changes to its background questionnaires (OECD, 2016a).

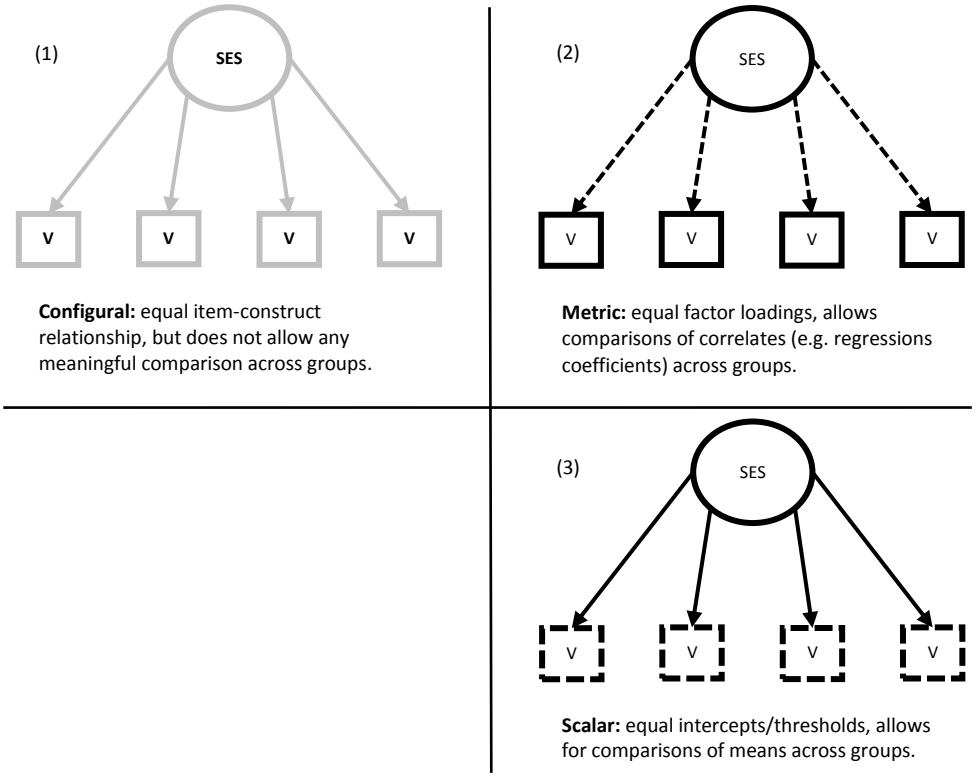
Analytical framework

Our analytical framework is broadly situated within measurement theory and more specifically within ideas of test theory and design (e.g., van der Linden, 2005; Wilson, 2005). Test theory is focused on how a set of observed responses can map onto a theoretical, unobservable construct. Within ILSAs these observed responses are elicited from a (standardized) test or instrument, which may be defined as "a technique of relating something we observed in the real world (sometimes called manifest or observed) to something we are measuring that only exists as part of a theory (sometimes called latent or unobserved)" (Wilson, p. 4). Instrument design, or the process of developing items that elicit an unobserved theoretical construct, is an iterative process. An underlying assumption in test theory, which governs instrument design, is that the relationship between the theoretical construct and the observed responses to the items that make up the instrument is a causal one (Wilson, 2005). That is, a respondent's level on a particular construct (or constructs in the multidimensional case) causes their responses for a set of items. Because we cannot observe the construct directly, the causal agent is latent and the measure is left to "infer the underlying construct" allowing the researcher to only assume causality (Wilson, p. 12).

In order to assume a causal relationship between the observed responses and the latent trait, the instrument development requires a rigorous validation process (Cronbach and Meehl, 1955; Messick, 1984). In this case, a valid instrument is one where there is ample evidence that suggests the items are measuring the intended theoretical construct for the selected population. The accumulation of the evidence is sometimes referred to as the validation process (Shadish, Cook, and Campbell, 2002). Although the validation process includes multiple steps, one important aspect is to test the correlational structure of the instrument within the intended population of respondents. Assuming the factor structure holds for the given population, the instrument designer then takes additional steps toward validation, such as multitrait-multimethod studies.

When an instrument is intended to be used with multiple populations, as is intended with ILSAs, further validation is required to ensure the instrument operates in the same way across all populations. Brown (2015) outlined four types of group invariance for this purpose: a) equal form, b) equal loadings, c) equal intercepts/ thresholds, and d) equal residual variances (also known as configural, metric, scalar and strict invariance, respectively). The equal form is the most lenient type of invariance and means that the structure of the item-construct relationship is identical across all groups (see quadrant 1 in Figure 1 for a graphical representation). The test of equal loadings builds upon the previous structure and requires that the true score variance in each item is identical across all groups (see quadrant 2 in Figure 1). Next, equal intercepts for continuous items and equal thresholds for discrete items demonstrates that the items have the same locations in the latent space (see quadrant 3 in Figure 1). Finally, equal residual variances, when built upon equal intercepts/thresholds, indicates that all items have the same amount of variance across each group since the loading plus the residual variance equals the total variance. Ensuring measurement invariance indicates that the same construct is being measured in the same way across different groups. Evidence of measurement equivalence does not automatically validate the causal relationship between the construct and respondent; however, an inability to demonstrate equivalence across populations suggests that the assumption of causality between the respondent and construct does not hold. It is also important to mention that the level of invariance required depends on the objectives of the analysis. Different levels of invariance allow for different types of comparisons. See Figure 1 for a summary of the types of comparisons allowed by the different levels of invariance.

Figure 1. Different levels of invariance and types of comparisons allowed in each level



Note: The dotted lines represent the part of the model that is being tested in each invariance level

Methodology

Data

The data for this study has been sourced from the latest cycles of three major ILSAs: TERCE, managed by the UNESCO's Latin American Laboratory for Assessment of the Quality of Education (LLECE); TIMSS, managed by the International Association for the Evaluation of Educational Achievement (IEA), and PISA managed by the Organisation for Economic Cooperation and Development (OECD). All three studies, TERCE, TIMSS 2015 and PISA 2015 are the most recent international comparative studies that assess student achievement and gather information from a range of educational stakeholders. Specifically, PISA measures student achievement in mathematics, science and reading of 15 years' old students, with a focus on students' ability to apply knowledge in practical contexts and 'everyday life' situations (OECD, 2014, p. 24). In contrast to PISA's focus on practical situations, TIMSS and TERCE are curriculum-based tests and focus on what students had an opportunity to learn in school. TIMSS measures students' achievement in mathematics and science at 4th and 8th grade (Mullis, I.V.S., Martin, M.O., Foy, P., and Hooper, M., 2016), while TERCE measures reading, mathematics and science at 3rd and 6th grades (Treviño, et al., 2015). In this study, we used data from the 72 education systems participating in PISA 2015, from the 44 education systems participating TIMSS 8th grade, and from the 16 education systems that participated in TERCE 6th grade.

For each of the studies, we purposely selected a scale that each testing organisation has constructed and included in their released database as a proxy measure of family background. These scales result from the student questionnaire that is administered to each participant after they complete the cognitive portion of the assessment. In PISA, student's socio-economic status is estimated by the index of economic, social and cultural status (ESCS), which is derived from several variables related to students' family background: parents' education, parents' occupation, a number of home possessions that can be taken as proxies for material wealth and cultural possessions, and the number of books available in the home (OECD, 2016c, p.205). In TIMSS, we

used the Home Educational Resources Scale (HERS), which was created based on students' responses concerning the availability of three resources: number of books in the home, highest level of parental education, and number of home study supports (Martin, Mullis, Hooper, Yin, Foy and Palazzo, 2016). In TERCE we used the Family Socioeconomic and Cultural Status Scale (FSCS), which was derived from the following items: parental education, parental occupation, family income, and availability of different home possessions and services (UNESCO-OREALC, 2016). Although the theoretical constructs are not the same across studies, our analyses are intended to examine the extent to which testing organizations are able to create scales that are comparable among the countries that participate in their studies, rather than compare the same scale across studies. Table 1 shows the set of indicators used in each study to measure socioeconomic background.

Table 1. Indicators used in each study to construct a proxy measure of socioeconomic background

Scale / Study	Item	Description
PISA: Index of economic, social and cultural status (ESCS)[2]	1. Highest occupational status of parents (HISEI). 2. Highest educational level of parents (PARED). 3. Home possessions (HOMEPOS).	1. Occupational data for both the student's father and mother were obtained from responses to open-ended questions. The responses were coded to four-digit ISCO codes and then mapped to the international socio-economic index of occupational status. 2. Highest level of education od either parent, recoded onto the following categories: (0) none, (1) primary education, (2) lower secondary, (3), vocational/pre-vocational upper secondary, (4) general upper secondary and/or non-tertiary post-secondary, (5) vocational tertiary and (6) and/or theoretically oriented tertiary and post-graduate. The index corresponds to the higher ISCED level of either parent. 3. Students reported the availability of 16 household items at home, including three country-specific household items and the amount of books at home. Then a summary index of all household and possession items was calculated using IRT modelling with WLEs (logits) for the latent dimensions which were transformed to scales with an average of 0 and a standard deviation of 1 (with equally weighted samples).
TIMSS: Home Educational Resources (HERS)	1. Number of books in the home. (BSBG04). 2. Number of home study supports (BSDG06S). 3. Highest level of education of either parent (BSDGEDUP).	1. Response categories: (1) 0-10, (2) 11-25, (3) 26-100, (4) 101-200, (5) More than 200. 2. Response categories: (1) None, (2) Internet connection or own room, (3) Both. 3. Highest level of education of either parent, recoded onto the following categories: (1) finished primary or some lower-secondary or did not go to school, (2) finished lower-secondary, (3) finished upper-secondary, (4) finished post-secondary education, (5) finished university of higher.
TERCE: Family Socioeconomic and Cultural Status Scale (FSCS)	1. Highest level of education of the mother (DQFIT09_02). 2. Highest occupational level of the mother (DQFIT11_02). 3. Monthly household income (DQFIT12). 4. Material the floor is made of in the home (DQFIT14). 5. Services in the home (BIENES1). 6. Home possessions (BIENES2). 7. Number of books in the home (DQFIT21).	1. Response categories: (1) none, (2) primary education, (3) more than primary education 2. Response categories: (1) has never worked out of the household, (2) cleaning, maintenance, construction, farmer, etc. (3) sales, operated machines, driver, etc., (4) administrative, owner of smal business, (5) professional, owner of medium/large business, managerial, etc. 3. Income declared recoded into country-income deciles with the following categories: (1) decil 1, (2) decil 2, (3) decil 3, (4) decil 4, (5) decil 5, (6) decil 6 to 10. 4. Response categories: (1) dirt, (2) cement or non-polished wood, (3) tiles or similar, (4) carpet, parquet or polished wood. 5. Students reported the availability of 5 services in the home: drainage, garbage collection, telephone landline, cable TV and internet connection . Then a summary index of all service items was calculated using principal component analysis (PCA). 6. Students reported the amount of the following household items in the home: TV, .radio, PC, refrigerator, washing machine, smart phone, car Then a summary index of all home possessions was calculated using principal component analysis (PCA). 7. Response categories: (1) none, (2) 10 or less, (3) 11-20, (4) 21-30, (5) more than 31

Source: OECD, 2016c; Martin, et al., 2016; UNESCO-OREALC, 2016.

[2] These variables are in turn derived from a set of individual items. See (OECD, 2016c) for more details on the procedure followed to construct this scale.

Analytical strategy

Our analytical strategy consisted of two main steps. We first used confirmatory factor analysis (CFA) to test the factor structure of the model used in each study (i.e. TIMSS, PISA and TERCE) to measure some aspect socioeconomic background[3]. One CFA model was fit separately for each country in each study. Then, we used multi-group confirmatory factor analysis (MGCFA) to test for different levels of invariance across countries for each scale in each of the three studies described above. MGCFA (Jöreskog, 1971) is one of the most commonly used techniques to assess measurement invariance (Billiet, 2003). MGCFA is a straightforward extension of CFA that is used to evaluate group differences in means and covariances within a common factor model (Jöreskog, 1971); or as McGrath (2015) puts it, to evaluate overall model fit across multiple groups (education systems in our case).

In the first step, in order to evaluate the goodness of fit for each model in each country, we used four measures: the chi-squared test, comparative fit index (CFI), Tucker-Lewis index (TLI) and root mean square error of approximation (RMSEA). We followed the cut-off points proposed by Rutkowski and Svetina (2014) for the analyses in contexts where the number of groups is large and the sample sizes are large and varied (e.g. ILSA samples): $\leq .10$ for RMSEA, $\geq .95$ for CFI and TLI. Although the chi-square test is not considered to be useful in this context (Meade et al., 2008; Rutkowski and Svetina; 2014, Cheung and Rensvold, 2002), we also report chi-square statistics in order to analyse if the scales behave as expected across all conditions in that these values are generally larger as more constraints were placed on these models.

It is important to note that for those cases in which the socioeconomic scale is formed with only three indicators, such as in TIMSS and PISA, the one-factor solution is just-identified (i.e. does not have degrees of freedom). As a consequence, the evaluation of fit indexes is not possible because a three-indicator model has a perfect fit. In any case, according to Brown (2015) these "model[s] can still be evaluated in terms of the interpretability and strength of its parameter estimates (e.g., magnitude of factor loadings)" (pp. 71).

In the second step, in order to test for the invariance of the socioeconomic scales across groups (i.e. education systems), MGCFA models were fitted to all groups simultaneously within each study. That is, one MGCFA model was fit to all countries participating in TERCE, a second MGCFA model was fit to all countries participating in TIMSS, and a third MGCFA model was fit to all countries participating in PISA. According to the common practice in the field, we conducted a series of nested tests that proceed from least to most restrictive models. In this way, we started by testing each scale for configural invariance, followed by metric and scalar invariance. Although it is possible to test for strict invariance or equal residual variances (i.e. the fourth level of invariance) in the hierarchy proposed by Brown (2015), scalar invariance is sufficient for meaningful comparison of latent means across groups (Marsh et al., 2010; Meredith 1993).

We carried out two sets of analyses within this second step. First, in order to examine the performance of MGCFA fit measures, MGCFA models were fit to all countries simultaneously, by study, where the test of configural invariance was followed by the tests of metric and scalar invariance. Following Rutkowski and Svetina (2014), we term this first set of analyses as *overall fit measures.* We evaluate each model (i.e. configural, metric and scalar) using the same criteria presented above. That is, CFI and TLI should be no smaller than .95, and RMSEA should be no larger than .10.

Then, in order to test the plausibility of metric and scalar invariance, we use ΔCFI, ΔTLI, and ΔRMSEA between more and less restrictive models (configural vs. metric, and metric vs. scalar). We term this second set of analyses as *relative fit measures.* Considering the large and varying sample sizes and the relative high number of groups (i.e. educational systems), we use the approach proposed by Rutkowski and Svetina (2014). For the test to show metric invariance, these

[3] Even though the TIMSS scale is not a socioeconomic index, the Home Educational Resources Scale is commonly used in IEA's publications as measure to proxy student socioeconomic background. See for example: Martin et al., 2013; Erberber, et al., 2015; Trude and Gustafsson, 2016.

differences must be ΔCFI ≤ 0.020, ΔTLI ≤ 0.020, and ΔRMSEA 0.030. For the test to show scalar invariance, these differences must be ΔCFI ≤ 0.010, ΔTLI ≤ 0.010, and ΔRMSEA 0.010.

Results

First, we show the general results regarding the extent to which the empirical indicators correspond to the theoretical constructs proposed by each study, tested by CFA procedure for each study/scale and for each country. Second, we show the results of the multi-group analyses and the test for measurement invariance for each study/scale across the education systems participating in each study.

Step 1: Single-country analysis. We start our analysis with separate CFAs for each country in each study. Because the scales for TIMSS and PISA have only three items, there is one unique set of parameters that perfectly fit and reproduce the data (Harrington, 2009). For this reason, instead of presenting a table with the fit indexes (which would include only constant values), we follow the follow the approach proposed by Miranda and Castillo (2018) and present a graph showing the standardised factor loadings for each item. This allows us to evaluate the models in terms of the magnitude of the factor loadings of each item (Brown, 2015). Figure 1 shows the factor loadings for the PISA scale, Figure 2 for the TIMSS scale, and Figure 3 for the TERCE scale. Even when the model for TERCE is over-identified (df > 0), for consistency purposes, we present the graph with the standardised factor loadings. In Figures 1, 2 and 3, each dot represents the standardised factor loading of each item in one given country, and the horizontal line crossing each dot represents the confidence interval at the 95% level. We marked a vertical line at a 0.5 factor loading as this can be considered the minimum acceptable value for standardised loading in CFA (Hair et al., 2006).

Figure 2. Standardised factor loadings for each item composing SES in PISA, TIMSS and TERCE

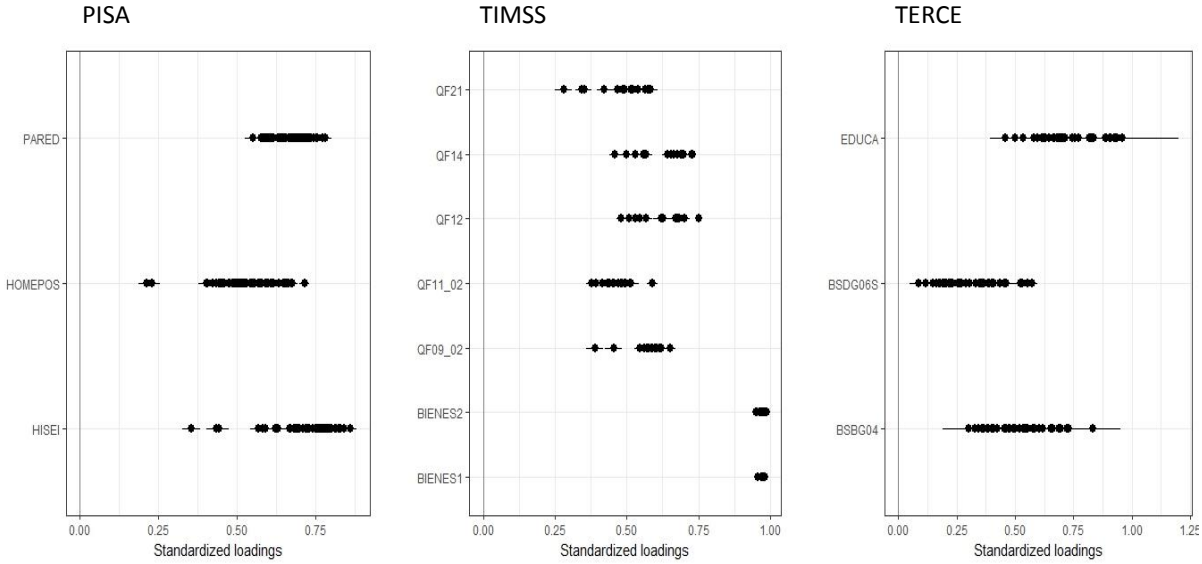


Table 2. Number of countries under/above 0.5 factor loading by indicator/study

Study	Indicator	Factor loading	
		< 0.5	**> 0.5**
PISA	HISEI	3	65
	PARED	0	68
	HOMPOS	25	43
TIMSS	BSBG04	16	28

TERCE	BSDG06S	37	7
	EDUCA	1	43
	QF09_02	0	16
	QF21	7	9
	QF11_02	12	4
	QF12	1	15
	QF14	1	15
	BIENES1	0	16
	BIENES2	0	16

Figure 2 presents the factor loadings of each indicator used for measuring socioeconomic background in the three studies, and Table 3 shows the number of countries in which the factor loadings of each indicator are below and above 0.5 for each of the studies considered. As can be observed, for PISA, the indicators PARED and HISEI show factor loadings above 0.5 values in most countries (68 and 65, respectively. See Table 2). The indicator HOMEPOS presents 25 countries with factor loadings below 0.5, and even two countries with values under 0.25 (see Table 2). For TIMSS, as can be observed in Figure 2, the indicator EDUCA is the only one that has factor loadings above 0.5 for most countries (43, see Table 2). The other two indicators present higher variations across countries. For instance, the indicator BSBG04 presents 16 countries with factor loadings under 0.5, and the indicator BSDG06S presents factor loadings under 0.5 in 37 countries, with about half of these countries with values under 0.25 (see Table 2). Finally, in TERCE, Figure 2 shows that none of the indicators presents factor loadings under 0.25. The items QF21 and QF11_02, however, present some countries with factor loadings under 0.5 (7 and 12, respectively. See Table 2). Particularly BIENES1 and BIENES2, present factor loadings above 0.5 in all the 16 countries participating in the study (see Table 2).

So far, we have illustrated that the analysed socioeconomic background measures and their configuration across countries show important variations among studies. Our results suggest that among the socioeconomic background scales analysed, the TERCE scales is the one with least variations in its configuration across countries and the one with the best fit, followed by the PISA and TIMSS scales.

Step 2: Multi-group analysis. The test of measurement invariance indicated different levels of invariance for the three analysed studies. In the case of PISA, using the information from the factor loadings, heuristically, it is reasonable to assume that the structure of the scale is similar across countries (see Figure 2). The three indicators had relatively stable estimates across countries, with factor loadings over 0.50 for most indicators in most countries. Only the HOMEPOS index showed some factor loadings under 0.25 (in Qatar and the United Emirates). As can be observed in Table 3, the metric model showed fit indices over the cut-off criteria, while the scalar model showed fit indices under the established criteria (see Table 3). However, the *relative fit* measures indicate that neither metric nor scalar invariance was achieved (see Table 4).

Table 3. MGCFA overall fit measures for each level of invariance.

	PISA				**TIMSS**				**TERCE**			
Model	X^2	*CFI*	*TLI*	*RMSEA*	X^2	*CFI*	*TLI*	*RMSEA*	X^2	*CFI*	*TLI*	*RMSEA*
Configural	0.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000	4124.600	0.979	0.968	0.070
Metric	5951.480	0.974	0.961	0.077	3195.150	0.863	0.790	0.077	6216.210	0.968	0.965	0.072
Scalar	61932.700	0.728	0.793	0.177	19723.990	0.146	0.672	0.096	16862.100	0.910	0.925	0.107

For TIMSS, the *overall fit* information (e.g. factor loadings) shows that baseline model of configural invariance indicates high dispersion in factor loadings (see Figure 2). Particularly, in only seven countries we found factor loadings for the 0.05, and about one-third of the countries (e.g. Canada, Hungary, Ireland, Italy, Japan, Kuwait) had factor loadings under 0.25. The other two indicators are relatively stable across countries. The metric model showed *overall* fit indices under the cut-off criteria, and as a consequence, there is no evidence to suggest that metric or scalar invariance was achieved (see Table 3). Similarly, the results of the *relative* fit indices suggest that neither metric nor scalar invariance was achieved (see Table 4).

Finally, the TERCE study showed good *overall* fit indices for the configural and metric models, but fit indices were out of the acceptable range for the scalar model (see Table 3 and Figure 2 for the factor loadings of the configural model). Regarding the *relative* fit measures, the comparison between the configural and metric models provide evidence of metric invariance (see Table 4).

In summary, our analyses resulted in fit indices that did not provide evidence of metric or scalar invariance for the socioeconomic background scales used in TIMSS and PISA, while the TERCE scale showed evidence of metric and scalar invariance.

Table 4. MGCFA relative fit measures for each level of invariance.

	PISA			**TIMSS**			**TERCE**		
Model	X^2*diff.*	Δ*CFI*	Δ*RMSEA*	X^2*diff.*	Δ*CFI*	Δ*RMSEA*	X^2*diff.*	Δ*CFI*	Δ*RMSEA*
Metric	5951.476	0.026	0.077	3126.102	-0.137	0.077	2086.991	-0.011	0.002
Scalar	44692.380	0.246	0.100	17830.425	-0.717	0.019	7523.642	-0.058	0.035

Discussion

At a basic level, background questionnaires are made of up of multiple instruments, parts of which are intended to measure a hypothetical construct (e.g., socioeconomic status). Because hypothetical constructs cannot be measured directly, answers to select background questionnaire items serve as an indirect indicator of the construct, operationalized through a measurement model. Because constructs are theoretical, unobservable phenomena, the act of verifying or validating the instrument is an extremely important process, a process that falls under modern test theory (e.g., van der Linden, 2005; Wilson 2005). The core idea of developing and validating such instruments is to begin with a well-defined construct, design a set of items that are assumed to elicit that construct, and then test if the proposed measurement model is consistent with the data generated from those items. When a proposed model does not fit data from one or more items, the items are revised or replaced. In other words, construct development should generally not be a posthoc exercise, where item responses are explored for viable constructs, which are then mapped back onto some theory. At least, it should not be a linear exercise that is finished with the best (but insufficient) attempt to fit the empirical data to a given theory.

Furthermore, only after sufficient evidence suggests the instrument is fitting well within a population can the task of evaluating the suitability of the instrument for cross-population comparisons begin. A common approach to testing measurement invariance across countries is to test the equality of the measurement model's covariance structure, means, and residual variances across countries. By examining the measurement model's degree of equality, we are able to statistically test the assumption of scale comparability. If the assumption holds then there would be statistical evidence that the scales can be compared sensibly. But, as was the case in the current examination, constructs are not always comparable suggesting that the traits being measured are not the same across cultures.

When scales are found to be non-invariant across populations or cultures a number of plausible explanations exist. First, and foremost, it is completely possible that a theorized construct is simply wrong or that it was once correct but changes in society have occurred so that the specified construct is no longer relevant. Of course, if the construct is not relevant the scale should not be reported and the old construct should be abandoned for a new construct. In situations where there exists strong theoretical backing for a universal construct there are other possible reasons that a scale may be found to be non-invariant to include:

1) The construct can be measured but the framework is incorrect;
2) The construct can be measured, the framework is correct, but the indicators are being operationalized incorrectly;
3) The construct can be measured, the framework is correct, but there are no universal indicators.

In terms of the first point, a viable and relatively straightforward treatment is that the framework used to operationalize the construct in question needs to be revised. The second point is a bit more nuanced. Operationalizing constructs incorrectly could happen for a variety of reasons. For example, the framework that stands as the foundation for measuring children's SES should include household income, which is a difficult question to reliably obtain from young children. Thus, other, more indirect indicators of household income must be collected. Potential alternatives could include the number of televisions, bedrooms, or books in the home. Although these might be the most reasonable variables to collect under the circumstance, measures of home possessions may not accurately reflect household income. For example, it is possible that with the dawn of e-readers the number of books in the home no longer represents either wealth or SES. In the third scenario, the construct exists but the indicators needed to measure that construct differ between countries or regions. Again, SES is a useful example to illustrate this point. The majority of academic theory may define SES as a universal construct. In support of the construct, a universal framework might be applied by researchers wishing to measure SES internationally. But regardless of the accepted

theory of SES, the indicators that represent the construct may differ by country. For example, a reliable indicator of family wealth in the U.S. might be whether the child has a room of their own or whether they have taken an international vacation. In contrast, a relatively poor indicator would be if the family has a lawnmower. In contrast, a lawnmower is a strong signal of wealth in Hong Kong or Singapore, given the relative lack of land on which to grow grass. The question of universal indicators but not a universal theory remains – is there a set of indicators that can reliably differentiate between well- and poorly-resourced children internationally? Clearly, given the performance of the measures examined here as well as similar research (Caro, Sandoval-Hernandez and Lüdtke, 2016; Rutkowski and Rutkowski, 2017) much work remains to be done.

One possible way forward is to relax the requirement that constructs should be identically defined across measured systems. Although PISA, as one example, has allowed for the inclusion of country-specific wealth items, these are not subsumed under a single latent variable model of SES. Rather, they are treated as observed, country-specific variables that are formed into linear combinations of variables to make up a measure of socio-cultural status. However, such linear combinations are not latent variables (Bentler, 1982), and have no hypothesized structure. Such approaches are not measuring anything but are merely exercises in data reduction.

It is possible to instead fit latent variable models that adhere to an assumption of partial invariance, whereby unique items and unique item parameters are allowed. Previous research has shown that, although more work needs to be done to operationalize this construct, it is a promising way to improve model-data consistency across countries while maintaining comparability. Importantly, Rutkowski and Rutkowski (2017) concentrated their efforts on the relatively homogeneous Nordic region and showed that more research needs to be done to develop well-functioning country-specific measures. Although this certainly will necessitate meaningful work on the part of participating countries, it will allow for participants to incorporate the local cultural nuances of their local context into internationally comparable scales.

TERCE provides another possible solution. As a regional assessment that focuses on similar language groups, cultures, and economies (when compared to PISA and TIMSS), with more focus TERCE should be able to design and administer questionnaires that are better tailored to a specific population. In the current manuscript, our results indicate that TERCE was able to develop a socioeconomic background scale that is comparable at the metric level which is better than its TIMSS counterpart. Regardless no study had acceptable scalar invariance where latent means can be validly compared across countries. In other words, for TERCE TIMSS and PISA, there is statistical evidence to suggest that the socio-economic background indicator is not cross-culturally comparable. Even worse, in both PISA and TIMSS the scales are not meeting basic quality standards within many participating educational systems. As such, analyses that use mean values of

the socioeconomic scales on any of these studies will produce findings that are questionable, at best. These findings have direct policy and research implications. For example, studies that estimate the share of resilient students[4] in a group of countries and then make cross-national comparisons are a classic example of such practice (e.g. OECD, 2011; Erberber, Stephens, Mamedova, Ferguson and Kroeger, 2015). Furthermore, the same can be said about any international comparative study that uses the PISA index of economic, social and cultural status (ESCS) or the TIMSS' home educational resources scale (HERS) as a control variable in a regression model[5]. Our findings pose a serious threat to the validity of these scales and any future analysis should caution readers to these threats.

Conclusion

Scales from the background questionnaire play an important role in helping to explain educational achievement. In fact, certain scales have taken a life of their own and often operate outside of the achievement results. For example, scales such as bullying, student engagement, and civic engagement are important to policy and interesting in absence of their relationship to achievement. Although, we used SES, a scale common to all three assessments, as an example for this study a similar analysis should be completed for any study that wishes to use scales from ILSAs for cross cultural comparison. Further, as demonstrated in this paper, substantial work needs to be done to improve measures internationally. At the very least, ILSA background scales require a validation process as rigorous as the achievement scales (OECD, 2014). Such a process would go a long way in preventing reporting scales that are not comparable across participating countries.

As hinted at by our results, by embracing a more rigours regional focus to questionnaire development, ILSAs might improve the comparability of certain constructs such as SES. More specifically, improving regional development of questionnaires or further funding the development of regional ILSAs and there questionnaires are two possible ways forward. It could be argued that the IEA's International Civic and Citizenship Study (ICCS), with its regional modules, represents the former and TERCE represents the latter of these possible models. In each case, however, a clear framework that maps directly onto regional specific scales is currently lacking and would need to be fully developed. In the case of larger ILSAs such as PISA and TIMMS, we recommend, at the very least, diversifying the cultural makeup of those stakeholders who oversee the current international frameworks. For example, the expert group that oversaw the PISA 2015 questionnaire framework and instruments committee included eleven members that were mostly from highly developed OECD economies (over half from the U.S. and Germany). The committee composition clearly did

[4] Commonly defined as students with low SES and high academic achievement.
[5] According to our results this type of comparison would be valid when using the TERCE's family socioeconomic and cultural status scale (FSCS) as this scale reached metric invariance (see tables 4 and 5).

not represent the extremely diverse cultural makeup of PISA participants (OECD, 2016b). Finally, in cases where countries or groups of countries find that the framework is miss-specified, members should work with the testing organizations to make adjustments to the framework and scales. If that is not possible, then participants should ask ILSA organizations not to include their country in any scale reporting. This engagement, of course, comes with a cost; however, publishing and using poor scales can be even more costly.

References

Bentler, P. M. (1982). Confirmatory factor analysis via non-iterative estimation. A fast inexpensive method. *Journal of Marketing Research*, *25A*(5), 309-318.

Billiet, J. (2003). Cross-cultural equivalence with structural equation modeling. En J. Harkness, F. Van de Vijver and P. Mohler (Eds.), *Cross-cultural survey methods (*pp. 247-264). NJ: John Wiley and Sons.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research.* New York: Guildford Press.

Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. En A. C. Porter and A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 150–197). Washington, DC: National Academy Press.

Caro, D. H., Sandoval-Hernández, A. and Lüdtke, O. (2016). Cultural, social, and economic capital constructs in international assessments: an evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, *25*(3), 433-450.

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. DOI: 10.1207/S15328007SEM0902_5

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Erberber, E., Stephens, M., Mamedova, S., Ferguson, S., and Kroeger, T. (2015). Socioeconomically disadvantaged students who are academically successful: Examining academic resilience cross-nationally. *IEA's Policy Brief Series*, No. 5, Amsterdam: IEA. Recuperado de http://www.iea.nl/policy_briefs.html

Gaviria Soto, J. L., Biencinto López, M. C., and Navarro Asencio, E (2007). Invarianza de la estructura de covarianzas de las medidas de rendimiento académico en estudios longitudinales en la transición de Educación Primaria a Secundaria. *Revista de Educación*, 348, 153-173.

Glas, C., and Jehangir, K. (2014). Modeling country-specific differential item functioning. En L. Rutkowski, M. von Davier, and D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis.* Boca Raton, FL: Chapman and Hall / CRC Press.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy and Practice*, *11*(3), 319–330. DOI: 10.1080/0969594042000304618

Harrington, D. (2009). *Confirmatory Factor Analysis*. Oxford: Oxford University Press.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 408-426.

Kreiner, S., and Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*(2), 210–231. DOI: 10.1007/s11336-013-9347-z

Martin, M.O., Mullis, I.V.S., Hooper, M., Yin, L., Foy, P. and Palazzo, L. (2016). Creating and Interpreting the TIMSS 2015 Context Questionnaire Scales. En M. O. Martin, I. V. S. Mullis, and M. Hooper (Eds.) *Methods and Procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center / IEA.

Marsh, H.W., Ludtke O., Muthen, B., Asparouhov, T, Morin, A.J.S., et al. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychol. Assess. 22*(3), 471–91.

McGrath, R. E. (2015). Measurement Invariance in Translations of the VIA Inventory of Strengths. *European Journal of Psychological Assessment*. On-line advanced publication. DOI: 10.1027/1015-5759/a000248

Mazzeo, J., and von Davier, M. (2009). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Education Working Papers EDU/PISA/GB.* Paris: OECD.

Messick, S. (1984). The Psychology of Educational Measurement. *Journal of Educational Measurement, 21*(3), 215–237.

Miranda, D. and Castillo, J. C. (2018). Measurement model and invariance testing of scales measuring egalitarian values in ICCS 2009. En A. Sandoval-Hernandez, M. M. Isac and D. Miranda (Eds.) *Teaching Tolerance in a Globalized World*. Cham: Springer International Publishing

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika 58*(4), 525–43

Mullis, I.V.S., Martin, M.O., Foy, P., and Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Chestnut Hill, MA: TIMSS and PIRLS International Study Center / IEA.

OECD. (2011). *Against the Odds: Disadvantaged Students who Succeed in School*. Paris: OECD Publishing.

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

OECD. (2014). *PISA 2012 Technical Report*. Paris: OECD Publishing.

OECD. (2016a). *PISA 2015 Assessment and Analytical Framework.* Paris: OECD Publishing.

OECD. (2016b). PISA 2015 Background questionnaires. Annex A (pp. 129–196). En *PISA 2015 Assessment and Analytical Framework*. Paris: OECD Publishing.

Oliveri, M. E., and Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education, 24*(4), 349–366. DOI: 10.1080/08957347.2011.607063

Oliveri, M. E., and von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315–333.

Oliveri, M. E., and von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*(1), 1–21. DOI: 10.1080/15305058.2013.825265

Rutkowski, L. and Rutkowski, D. (2010). Getting it "better": The importance of improving background questionnaires in International Large-Scale Assessment. *Journal of Curriculum Studies, 42*(3), 411–430. DOI: 10.1080/00220272.2010.487546

Rutkowski, L. and Rutkowski, D. (2017). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research, 0*(0), 1–14. DOI: 10.1080/00313831.2016.1261044

Rutkowski, L., Rutkowski, D. and Zhou, Y. (2016). Parameter estimation methods and the stability of achievement estimates and system rankings: Another look at the PISA model. *International Journal of Testing, 16*(1), 1–20.

Rutkowski, L., y Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31-57.

Schulz, W., Ainley, J. and Fraillon, J. (Eds.). (2011). *ICCS 2009 Technical Report*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Treviño, E., Fraser, P., Meyer, A., Morawietz, L., Inostroza, P. and Naranjo, E. (2015). *Informe de Resultados TERCE. Factores Asociados*. Santiago: Oficina Regional de Educación para América Latina y el Caribe (OREAL/UNESCO).

UNESCO-OREALC. (2016). *Reporte Técnico. Tercer Estudio Regional Comparativo y Explicativo, TERCE*. Santiago: Oficina Regional de Educación para América Latina y el Caribe (OREAL/UNESCO).

Van Der Linden, W. J. (2005). A Comparison of Item-Selection Methods for Adaptive Tests with Content Constraints. *Journal of Educational Measurement*, *42*(3), 283-302.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Lawrence Erlbaum Associates.